

# Lower Bound on the Redundancy of PIR Codes

Sankeerth Rao and Alexander Vardy

## Abstract

We prove that the redundancy of a  $k$ -server PIR code of dimension  $s$  is  $\Omega(\sqrt{s})$  for all  $k \geq 3$ . This coincides with a known upper bound of  $O(\sqrt{s})$  on the redundancy of PIR codes. Moreover, for  $k = 3$  and  $k = 4$ , we determine the lowest possible redundancy of  $k$ -server PIR codes exactly. Similar results were proved independently by Mary Wootters using a different method.

Given two binary vectors  $\mathbf{u} = (u_1, u_2, \dots, u_n)$  and  $\mathbf{v} = (v_1, v_2, \dots, v_n)$ , we define their *product*  $\mathbf{uv}$  componentwise, namely

$$\mathbf{uv} \stackrel{\text{def}}{=} (u_1v_1, u_2v_2, \dots, u_nv_n) \quad (1)$$

where  $u_1v_1, u_2v_2, \dots, u_nv_n$  are computed in  $\text{GF}(2)$ . Note that the product operation in (1) distributes over addition in  $\mathbb{F}_2^n$ . Thus (1) turns the vector space  $\mathbb{F}_2^n$  into an algebra  $\mathcal{A}_n$  over  $\mathbb{F}_2$ . This algebra  $\mathcal{A}_n$  is unital, associative, and commutative.

Given a set  $X \subseteq \mathbb{F}_2^n$ , we define the square of  $X$  as the set of products of the elements in  $X$ . Explicitly,  $X^2$  is defined as follows:

$$X^2 \stackrel{\text{def}}{=} \{ \mathbf{uv} : \mathbf{u}, \mathbf{v} \in X \text{ and } \mathbf{u} \neq \mathbf{v} \} \quad (2)$$

The following lemmas follow straightforwardly from the definitions in (1) and (2), along with the fact that  $\mathcal{A}_n$  is a commutative algebra. We let  $\langle X \rangle$  denote the linear span over  $\mathbb{F}_2$  of a set  $X \subseteq \mathbb{F}_2^n$ .

**Lemma 1.**  $|X^2| \leq |X|(|X| - 1)/2$ .

*Proof.* If  $|X| = r$ , then  $X^2$  consists of the  $\binom{r}{2}$  vectors  $\mathbf{uv} = \mathbf{vu}$  for some  $\mathbf{u} \neq \mathbf{v}$  in  $X$ . Some of these vectors may coincide.  $\square$

**Lemma 2.** Let  $\mathbf{u}, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in \mathbb{F}_2^n$ . If  $\mathbf{v}_1\mathbf{v}_2 + \mathbf{v}_1\mathbf{v}_3 + \mathbf{v}_2\mathbf{v}_3 = \mathbf{0}$ , then

$$(\mathbf{u} + \mathbf{v}_1)(\mathbf{u} + \mathbf{v}_2) + (\mathbf{u} + \mathbf{v}_2)(\mathbf{u} + \mathbf{v}_3) + (\mathbf{u} + \mathbf{v}_3)(\mathbf{u} + \mathbf{v}_1) = \mathbf{u}$$

*Proof.* Follows by straightforward verification using distributivity and commutativity in  $\mathcal{A}_n$ .  $\square$

We now show how the foregoing lemmas can be used to establish a bound on the redundancy of binary  $k$ -server PIR codes for  $k \geq 3$ . These codes are defined in [1,2] as follows.

**Definition 1.** Let  $\mathbf{e}_i$  denote the binary (column) vector with 1 in position  $i$  and zeros elsewhere. We say that an  $s \times n$  binary matrix  $G$  has **property**  $\mathcal{P}_k$  if for all  $i \in [s]$ , there exist  $k$  disjoint sets of columns of  $G$  that add up to  $\mathbf{e}_i$ . A matrix that has property  $\mathcal{P}_k$  is also said to be a  **$k$ -server PIR matrix**. A binary linear code  $\mathcal{C}$  of length  $n$  and dimension  $s$  is called a  **$k$ -server PIR code** if there exists a generator matrix  $G$  for  $\mathcal{C}$  with property  $\mathcal{P}_k$ .

For much more on  $k$ -server PIR codes and their applications in reducing the storage overhead of private information retrieval, see [1,2]. In particular, it is shown in [2] that, given a  $k$ -server PIR code of length  $s + r$  and dimension  $s$ , the storage overhead of *any* linear  $k$ -server PIR protocol can be reduced from  $k$  to  $(s + r)/s$ . Moreover, for every fixed  $k$ , there exist  $k$ -server PIR codes whose rate (and, hence, storage overhead) approaches 1 as their dimension  $s$  grows. However, exactly *how fast* the resulting storage overhead tends to 1 as  $s \rightarrow \infty$  was heretofore unknown. For every fixed  $k$ , Fazeli, Vardy, and Yaakobi [1,2] construct  $k$ -server PIR codes with redundancy  $r$  bounded by  $r \leq k\sqrt{s}(1 + o(1))$ . But the question of whether codes with even smaller redundancy exist was left open in [1,2]. The following theorem shows that the redundancy  $O(\sqrt{s})$  of the codes constructed in [1,2] is asymptotically optimal.

**Theorem 3.** *Let  $\mathbb{C}$  be a 3-server PIR code of length  $n$  and dimension  $s$ . Let  $r = n - s$  denote the redundancy of  $\mathbb{C}$ . Then  $r(r - 1) \geq 2s$ .*

*Proof.* Let  $G$  be an  $s \times n$  generator matrix for  $\mathbb{C}$  with property  $\mathcal{P}_3$ , and let  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  denote the columns of  $G$ . By definition, for each  $i \in [s]$ , there exist 3 disjoint subsets of  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  that add up to  $\mathbf{e}_i$ . Let  $R_1, R_2, R_3 \subset [n]$  denote the corresponding sets of indices. Then we can write

$$\mathbf{e}_i = \sum_{j \in R_1} \mathbf{x}_j = \sum_{j \in R_2} \mathbf{x}_j = \sum_{j \in R_3} \mathbf{x}_j \quad (3)$$

It is easy to see from Definition 1 that  $G$  has full column rank. Hence some  $s$  columns of  $G$  are linearly independent, and we assume w.l.o.g. that these are the first  $s$  columns. Consequently, there exists a nonsingular  $s \times s$  matrix  $A$  such that

$$G' \stackrel{\text{def}}{=} AG = [I_s | P] \quad (4)$$

where  $I_s$  is the  $s \times s$  identity matrix and  $P$  is an  $s \times r$  matrix. Let  $\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_n$  denote the columns of  $G'$ , with  $\mathbf{x}'_j = \mathbf{e}_j$  for  $j = 1, 2, \dots, s$ . Then it follows from (3) that

$$\mathbf{a}_i = \sum_{j \in R_1} \mathbf{x}'_j = \sum_{j \in R_2} \mathbf{x}'_j = \sum_{j \in R_3} \mathbf{x}'_j \quad (5)$$

where  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_s$  are the columns of  $A$ . Note that  $\dim \langle \mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_s \rangle = s$ , since the matrix  $A$  is nonsingular. Let us now further define

$$S_1 = R_1 \cap [s], \quad S_2 = R_2 \cap [s], \quad S_3 = R_3 \cap [s] \quad (6)$$

$$T_1 = R_1 \cap ([n] \setminus [s]), \quad T_2 = R_2 \cap ([n] \setminus [s]), \quad T_3 = R_3 \cap ([n] \setminus [s]) \quad (7)$$

$$\mathbf{v}_1 = \sum_{j \in S_1} \mathbf{x}'_j = \sum_{j \in S_1} \mathbf{e}_j \quad \mathbf{v}_2 = \sum_{j \in S_2} \mathbf{x}'_j = \sum_{j \in S_2} \mathbf{e}_j \quad \mathbf{v}_3 = \sum_{j \in S_3} \mathbf{x}'_j = \sum_{j \in S_3} \mathbf{e}_j \quad (8)$$

With this notation, we can rewrite (5) as follows:

$$\mathbf{a}_i + \mathbf{v}_1 = \sum_{j \in T_1} \mathbf{x}'_j \quad \mathbf{a}_i + \mathbf{v}_2 = \sum_{j \in T_2} \mathbf{x}'_j \quad \mathbf{a}_i + \mathbf{v}_3 = \sum_{j \in T_3} \mathbf{x}'_j \quad (9)$$

Finally, let us define  $X \stackrel{\text{def}}{=} \{\mathbf{x}'_{s+1}, \mathbf{x}'_{s+2}, \dots, \mathbf{x}'_n\}$ . Then it follows from (9) that  $\mathbf{a}_i + \mathbf{v}_1, \mathbf{a}_i + \mathbf{v}_2$ , and  $\mathbf{a}_i + \mathbf{v}_3$  belong to  $\langle X \rangle$ . We are now ready to use Lemmas 1 and 2 in order to complete the proof.

Since the sets  $S_1, S_2, S_3$  are disjoint, it follows from (8) that the supports of  $v_1, v_2, v_3$  are also disjoint. In other words,  $v_1 v_2 = v_1 v_3 = v_2 v_3 = \mathbf{0}$ . Using Lemma 2, we conclude that

$$\begin{aligned} a_i &= (a_i + v_1)(a_i + v_2) + (a_i + v_2)(a_i + v_3) + (a_i + v_3)(a_i + v_1) \\ &= \left( \sum_{j \in T_1} x'_j \right) \left( \sum_{j \in T_2} x'_j \right) + \left( \sum_{j \in T_2} x'_j \right) \left( \sum_{j \in T_3} x'_j \right) + \left( \sum_{j \in T_3} x'_j \right) \left( \sum_{j \in T_1} x'_j \right) \\ &= \sum_{j \in T_1} \sum_{k \in T_2} x'_j x'_k + \sum_{j \in T_2} \sum_{k \in T_3} x'_j x'_k + \sum_{j \in T_3} \sum_{k \in T_1} x'_j x'_k \end{aligned}$$

Since the sets  $T_1, T_2, T_3$  are disjoint subsets of  $[n] \setminus [s]$ , all of the products  $x'_j x'_k$  above belong to  $X^2$ . Consequently, it follows that  $a_i \in \langle X^2 \rangle$  for all  $i$ . Hence

$$\dim \langle X^2 \rangle \geq \dim \langle a_1, a_2, \dots, a_s \rangle = s$$

But  $\dim \langle X^2 \rangle \leq |X^2| \leq r(r-1)/2$ , where we have used Lemma 1. Thus  $r(r-1)/2 \geq s$ , which completes the proof of the theorem.  $\square$

It is shown in [1,2] that the redundancy of  $k$ -server PIR codes is non-decreasing in  $k$ . That is, if  $\rho(s, k)$  denotes the lowest possible redundancy of a  $k$ -server PIR code of dimension  $s$ , then

$$\rho(s, k+1) \geq \rho(s, k) \quad \text{for all } s \geq 1 \text{ and all } k \geq 2$$

Consequently, the lower bound of Theorem 3 trivially extends from 3-server PIR codes to general  $k$ -server PIR codes with  $k \geq 3$ .

The following simple construction achieves the lower bound of Theorem 3 for  $k = 3$ . Let  $r$  be the smallest integer such that  $\binom{r}{2} \geq s$ . Take  $G = [I_s \mid P]$ , where  $P$  is an  $s \times r$  matrix whose rows are distinct binary vectors of weight 2. Clearly, the rows of  $P$  form a constant-weight binary code with distance 2. By the results of [1,2], this implies that  $G$  is a 3-server PIR matrix, and therefore

$$\rho(s, 3) = \text{the smallest integer } r \text{ such that } r(r-1) \geq 2s = \left\lceil \sqrt{2s + \frac{1}{4}} + \frac{1}{2} \right\rceil \quad (10)$$

It is also shown in [1,2] that for all even  $k$ , we have  $\rho(s, k) = \rho(s, k-1) + 1$ . Consequently, (10) determines the lowest possible redundancy of 4-server PIR codes as well.

## References

- [1] A. FAZELI, A. VARDY, and E. YAAKOBI, Codes for distributed PIR with low storage overhead, *Proc. IEEE Symp. Information Theory (ISIT)*, pp. 2852–2856, Hong Kong, June 2015.
- [2] A. FAZELI, A. VARDY, and E. YAAKOBI, PIR with low storage overhead: Coding instead of replication, available online at [arXiv:1402.2011v1](https://arxiv.org/abs/1402.2011), May 2015.
- [3] M. WOOTTERS, Linear codes with disjoint repair groups, unpublished manuscript, February 26, 2016.